Celi: EDITS and Generic Text Pair Classification

Milen Kouylekov
Celi S.R.L.
via San Quintino 31
Torino,10121, Italy
kouylekov@celi.it

Luca Dini
Celi S.R.L.
via San Quintino 31
Torino,10121, Italy
dini@celi.it

Alessio Bosca
Celi S.R.L.
via San Quintino 31
Torino,10121, Italy
alessio.bosca@celi.it

Marco Trevisan
Celi S.R.L.
via San Quintino 31
Torino, Italy
trevisan@celi.it

Abstract

This paper presents CELI's participation in the SemEval The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge (Task7) and Cross-lingual Textual Entailment for Content Synchronization task (Task 8).

1 Introduction

Recognizing an existing relation between two text fragments received a significant interest as NLP task in the recent years. A lot of the approaches were focused in the filed of Textual Entailment(TE). TE has been proposed as as a comprehensive framework for applied semantics (Dagan and Glickman, 2004), where the need for an explicit mapping between linguistic objects can be, at least partially, bypassed through the definition of semantic inferences at the textual level. In the TE framework, a text (T) is said to entail the hypothesis (H) if the meaning of H can be derived from the meaning of T. Initially defined as binary relation between texts (**YES/NO** there is an entailment or there is not) the TE evolved in the third RTE3 (Giampiccolo et al., 2007) challenge into a set of three relations between texts: **ENTAILMENT**, **CONTRADICTION** and **UNKNOWN**. These relations are interpreted as follows:

- ENTAILMENT - The T entails the H.
- CONTRADICTION - The H contradicts the T
- UNKNOWN - There is no semantic connection between T and H.

With more and more applications available for recognizing textual entailment the researches focused their efforts in finding practical applications for the developed systems. Thus the Cross-Lingual Textual Entailment task (CLTE) was created using textual entailment (TE) to define cross-lingual content synchronization scenario proposed in (Mehdad et. al., 2011), (Negri et. al., 2011) (Negri et. al., 2012). The task is defined by the organizers as follows: Given a pair of topically related text fragments (T1 and T2) in different languages, the CLTE task consists of automatically annotating it with one of the following entailment judgments:

- Bidirectional: the two fragments entail each other (semantic equivalence)
- Forward: unidirectional entailment from T1 to T2
- Backward: unidirectional entailment from T2 to T1
- No Entailment: there is no entailment between T1 and T2

The textual entailment competition also evolved. In this year SEMEVAL The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge - JRSA-RTE8 (Task7) the textual entailment was defined in three subtasks:

5-way task , where the system is required to classify the student answer according to one of the following judgments:

- Correct, if the student answer is a complete and correct paraphrase of the reference answer;

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 592–597, Atlanta, Georgia, June 14-15, 2013.

- Partially_correct_incomplete, if the student answer is a partially correct answer containing some but not all information from the reference answer;
- Contradictory, if the student answer explicitly contradicts the reference answer;
- Irrelevant, if the student answer is "irrelevant", talking about domain content but not providing the necessary information;
- Non_domain, if the student answer expresses a request for help, frustration or lack of domain knowledge - e.g., "I don't know", "as the book says", "you are stupid".

3-way task , where the system is required to classify the student answer according to one of the following judgments:

- correct
- contradictory
- incorrect, conflating the categories of partially_correct_incomplete, irrelevant or non_domain in the 5-way classification

2-way task , where the system is required to classify the student answer according to one of the following judgments:

- correct
- incorrect, conflating the categories of contradictory and incorrect in the 3-way classification.

Following the overall trend, we have decided to convert our system for recognizing textual entailment EDITS from a simple YES/NO recognition system into a generic system capable of recognizing multiple semantic relationships between two texts.

EDITS (Kouylekov and Negri, 2010) and (Kouylekov et. al., 2011) is an open source package for recognizing textual entailment, which offers a modular, flexible, and adaptable working environment to experiment with the RTE task over different datasets. The package allows to: *i)* create an entailment engine by defining its basic components *ii)* train such entailment engine over an annotated RTE corpus to learn a model; and *iii)* use the entailment engine and the model to assign an entailment judgments and a confidence score to each pair of an unannotated test corpus.

We define the recognition of semantic relations between two texts as a classification task. In this task the system takes as an input two texts and classifies them in one of a set of predefined relations. We have modified EDITS in order to handle the so defined task.

Having this in mind we have participated in JRSA-RTE8 (task 7) and CLTE2 (task 8) with the same approach. We have merged EDITS with some features from the TLike system described in our last participation in CLTE (Kouylekov et. al., 2011). For each of the tasks we have created a specialized components that are integrated in EDITS as one of the system's modules.

2 EDITS and Generic Text Pair Classification

As in the previous versions, the core of EDITS implements a distance-based framework. Within this framework the system implements and harmonizes different approaches to distance computation between texts, providing both *edit distance* algorithms, and *similarity* algorithms. Each algorithm returns a normalized distance score (a number between 0 and 1). Each algorithm algorithm depends on two generic modules defined by the system's user:

- Matcher - a module that is used to align text fragments. This module uses semantic techniques and entailment rules to find equivalent textfragments.
- Weight Calculator - a module that is used to give weight to text fragments. The weights are used to determine the importance of a text portion to the overall meaning of the text.

In the previous versions of the system at the training stage, distance scores calculated over annotated T-H pairs are used to estimate a threshold that best separates positive (YES) from negative (NO) examples. The calculated threshold was used at a test stage to assign an entailment judgment and a confidence score to each test pair. In the new version

of the system we used a machine learning classifier to classify the T-H pairs in the appropriate category. The overall architecture of the system is shown in Figure 1.

The new architecture is divided in two sets of modules: Machine Learning and Edit Distance. In the Edit Distance set various distance algorithms are used to calculate the distance between the two texts. Each of these algorithms have a custom matcher and weight calculator. The distances calculated by each of these algorithms are used as features for the classifiers of the Machine Leaning modules. The machine learning modules are structured in two levels:

- Binary Classifiers - for each semantic relation we create a binary classifier that distinguishes between the members of the relation and the members of the other relations. For example: For 3way task (Task 7) the system created 3 binary classifiers one for each relation.
- Classifier - a module that makes final decision for the text pair taking the output (decision and confidence) of the binary classifiers as an input.

We have experimented with other configurations of the machine leaning modules and selected this one as the best performing on the available datasets of the previous RTE competitions. In the version of EDITS avalble online other configurations of the machine leaning modules will be available using the flexibility of the system configuration.

We have used the algorithms implemented in WEKA (Hall et al., 2009) for the classification modules. The binary modules use SMO algorithm. The top classifier uses NaiveBayes.

The input to the system is a corpus of text pairs each classified with one semantic relation. We have used the format of the previous RTE competitions in order to be compliant. The goal of the system is to create classifier that is capable of recognizing the correct relation for an un-annotated pair of texts.

The new version of EDITS package allows to:

- Create an *Classifier* by defining its basic components (*i.e.* algorithms, matchers, and weight calculators);
- Train such *Classifier* over an annotated corpus (containing T-H pairs annotated in terms of entailment) to learn a *Model*;
- Use the *Classifier* and the *Model* to assign an entailment judgment and a confidence score to each pair of an un-annotated test corpus.

3 Resources

Like our participation in the 2012 SemEval Cross-lingual Textual Entailment for Content Synchronization task (Kouylekov et. al., 2011), our approach is based on four main resources:

- A system for Natural Language Processing able to perform for each relevant language basic tasks such as part of speech disambiguation, lemmatization and named entity recognition.
- A set of word based bilingual translation modules.(Employed only for Task 8)
- A semantic component able to associate a semantic vectorial representation to words.
- We use Wikipedia as multilingual corpus.

NLP modules are described in (Bosca and Dini, 2008), and will be no further detailed here.

Word-based translation modules are composed by a bilingual lexicon look-up component coupled with a vector based translation filter, such as the one described in (Curtoni and Dini, 2008). In the context of the present experiments, such a filters has been deactivated, which means that for any input word the component will return the set of all possible translations. For unavailable pairs, we make use of triangular translation (Kraaij, 2003).

As for the semantic component we experimented with a corpus-based distributional approach capable of detecting the interrelation between different terms in a corpus; the strategy we adopted is similar to Latent Semantic Analysis (Deerwester et. al., 1990) although it uses a less expensive computational solution based on the Random Projection algorithm (Lin et. al., 2003) and (Bingham et. al., 2001). Different works debate on similar issues: (Turney, 2001) uses LSA in order to solve synonymy detection questions from the well-known TOEFL test while the method presented by (Inkpen, 2001) or by (Baroni and Bisi, 2001) proposes the use of the Web as a corpus to

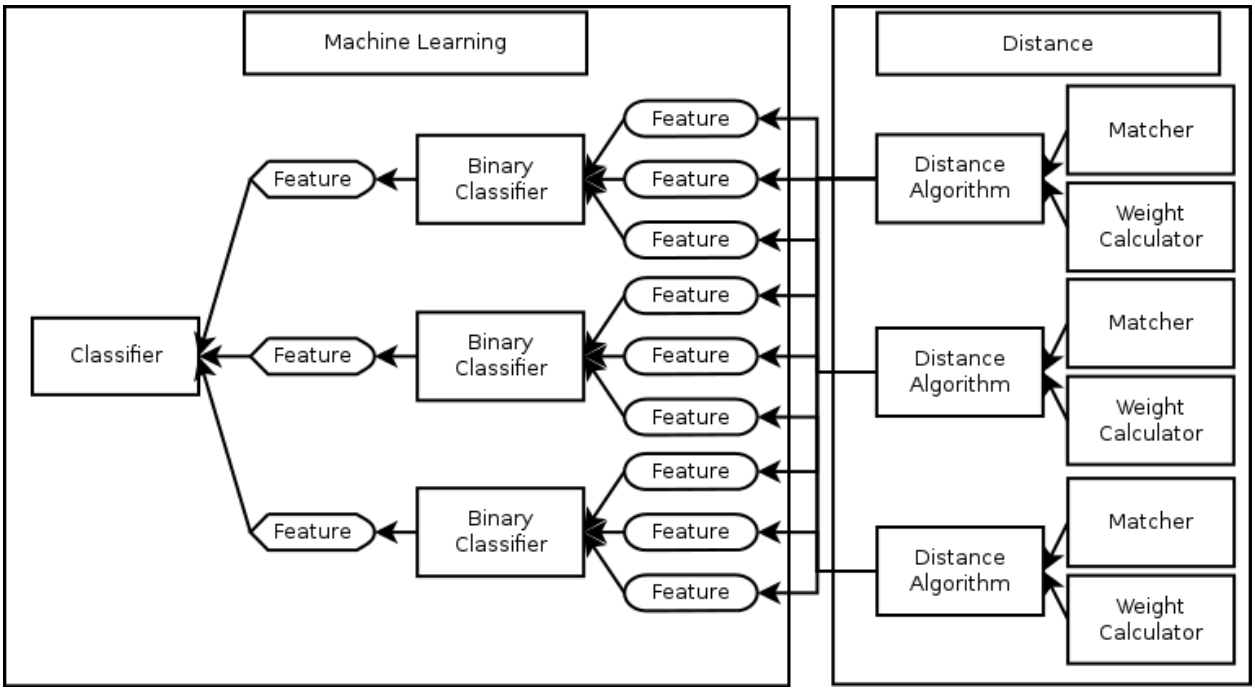

Figure 1: EDITS Architecture

compute mutual information scores between candidate terms.

We use Wikipedia as a corpus for calculating word statistics in different languages. We have indexed using Lucene[1] the English, Italian, French, German, Spanish distributions of the resource.

The semantic component and the translation[2] modules are used as core components in the matcher module. IDF calculated on Wikipedia is used as weight for the words by the weight calculator model.

4 JRSA-RTE8

In the JRSA-RTE8 we consider the reference answers as T (text) and the student answer as H (hypothesis). As the reference answers are often more than one, we considered as input to the machine learning algorithms the distance between the student answer and the **closest** reference answer. We define the closest reference answer as the reference answer with minimum distance according to the distance algorithm.

[1] *http://lucene.apache.org*

[2] Translation module is used only for Task 8.

4.1 Systems

We have submitted **two** runs in the SemEval JRSA-RTE8 challenge (Task 7). The systems were executed on each of the sub tasks of the main task.

System 1 The distance algorithm used in the first system is Word Overlap. The algorithm tries to find the words of a source text between the words of the target text. We have created two features for each binary classifier: 1) Feature 1 - word overlap of H into T (words of H are matched by the words in T; 2) Feature 2 - word overlap T into H (Words of T are matched by the words in H).

System 2 In the second system the we have used only Feature 1.

We have created separate models for the Beatle dataset and the sciEntsBank dataset. The results obtained are shown in Table 1.

4.2 Analysis

The results obtained are in line with our previous participations in the RTE challenges (Kouylekov et. al., 2011). Of course as we described before in our papers (Kouylekov et. al., 2011) the potential of the edit distance algorithm is limited. Still it provides a

Task	Beatle Q	Beatle A	sciEntsBank Q	sciEntsBank A	sciEntsBank D
2way					
run 1	0.6400	0.6570	0.5930	0.6280	0.6160
run 2	0.4620	0.4480	0.5560	0.5930	0.5710
3way					
run 1	0.5510	0.4950	0.5240	0.5780	0.5490
run 2	0.4150	0.4400	0.4390	0.5030	0.4770
5way					
run 1	0.4830	0.4470	0.4130	0.4340	0.4170
run 2	0.3850	0.4320	0.2330	0.2370	0.2540

Table 1: Task 7 Results obtained. (Accuracy)

good performance and provides a solid potential for some close domain tasks as described in (Negri and Kouylekov, 2009). We were quite content with the new machine learning based core. The selected configuration performed in an acceptable manner. The results obtained were in line with the cross accuracy obtained by our system on the training set which shows that it is not susceptible to over-training.

5 CLTE

5.1 Systems

We have submitted **two** runs in the CLTE task (Task 8).

System 1 The distance algorithm used in the first system is Word Overlap as we did for task 7. We have created two features for each binary classifier: 1) Feature 1 - word overlap of H into T (words of H are matched by the words in T; 2) Feature 2 - word overlap T into H (Words of T are matched by the words in H).

System 2 In the second system we have made a slight modification of the matcher that handled numbers.

The matcher module for this task used the translation modules defined in Section 3. We have created a model for each language pair.

The results obtained are shown in Table 2.

5.2 Analysis

The results obtained are quite disappointing. Our system obtained on the test set of the last CLTE competition (CLTE1) quite satisfactory results (clte1-test). All the results obtained for this competition are near or above the medium of the best systems. Our algorithm did not show signs of over-training (the accuracy of the system on the test and on the training of CLTE1 were almost equal). Having this in mind we expected to obtain scores at least in the margins of 0.45 to 0.5. This does not happen according us due to the fact that this year dataset has characteristics quite different than the last year. To test this hypothesis we have trained our system on half of the dataset (clte2-half-training) ,given for test this year, and test it on the rest (clte-half-test). The results obtained demonstrate that the dataset given is more difficult for our system than the last years one. The results also prove that our system is probably too conservative when learning from examples. If the test set is similar to the training it performs in consistent manner on both, otherwise it demonstrates severe over-training problems.

6 Conclusions

In this paper we have presented a generic system for text pair classification. This system was evaluated on task 7 and task 8 of Semeval 2013 and obtained satisfactory results. The new machine learning module of the system needs improvement and we plan to focus our future efforts in it.

We plan to release the newly developed system as version 4 of the open source package EDITS available at *http://edits.sf.net*.

Acknowledgments

This work has been partially supported by the ECfunded project Galateas (CIP-ICT PSP-2009-3-250430).

Run	Spanish	Italian	French	German
run1	0.34	0.324	0.346	0.349
run2	0.342	0.324	0.34	0.349
clte2-half-training	0.41	0.43	0.40	0.44
clte2-half-test	0.43	0.44	0.41	0.43
clte1-test	0.52	0.51	0.54	0.55

Table 2: Task 8. Results obtained. (Accuracy)

References

Baroni M., Bisi S. 2004. Using cooccurrence statistics and the web to discover synonyms in technical language In Proceedings of LREC 2004

Bentivogli L., Clark P., Dagan I., Dang H, Giampiccolo D. 2011. The Seventh PASCAL Recognizing Textual Entailment Challenge In Proceedings of TAC 2011

Bingham E., Mannila H. 2001. Random projection in dimensionality reduction: Applications to image and text data. In Knowledge Discovery and Data Mining, ACM Press pages 245250

Bosca A., Dini L. 2008. Query expansion via library classification system. In CLEF 2008. Springer Verlag, LNCS

Curtoni P., Dini L. 2006. Celi participation at clef 2006 Cross language delegated search. In CLEF2006 Working notes.

Dagan I. and Glickman O. 2004. Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. Learning Methods for Text Understanding and Mining Workshop.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41 391407

Giampiccolo; Bernardo Magnini; Ido Dagan; Bill Dolan. 2007. The Third PASCAL Recognizing Textual Entailment Challenge. Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. June 2007, Prague, Czech Republic

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. 2009 The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Inkpen D. 2007. A statistical model for near-synonym choice. ACM Trans. Speech Language Processing 4(1)

Kouylekov M., Negri M. An Open-Source Package for Recognizing Textual Entailment. 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010) ,Uppsala, Sweden. July 11-16, 2010

Kouylekov M., Bosca A., Dini L. 2011. EDITS 3.0 at RTE-7. Proceedings of the Seventh Recognizing Textual Entailment Challenge (2011).

Kouylekov M., Bosca A., Dini L., Trevisan M. 2012. CELI: An Experiment with Cross Language Textual Entailment. In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012).

Kouylekov M., Mehdad Y. and Negri M. 2011 Is it Worth Submitting this Run? Assess your RTE System with a Good Sparring Partner Proceedings of the TextInfer 2011 Workshop on Textual Entailment

Kraaij W. 2003. Exploring transitive translation methods. In Vries, A.P.D., ed.: Proceedings of DIR 2003.

Lin J., Gunopulos D. 2003. Dimensionality reduction by random projection and latent semantic indexing. In proceedings of the Text Mining Workshop, at the 3rd SIAM International Conference on Data Mining.

Mehdad Y.,Negri M., Federico M.. 2011. Using Parallel Corpora for Cross-lingual Textual Entailment. In Proceedings of ACL-HLT 2011.

Negri M., Bentivogli L., Mehdad Y., Giampiccolo D., Marchetti A. 2011. Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora. In Proceedings of EMNLP 2011.

Negri M., Kouylekov M., 2009 Question Answering over Structured Data: an Entailment-Based Approach to Question Analysis. RANLP 2009 - Recent Advances in Natural Language Processing, 2009 Borovets, Bulgaria

Negri M., Marchetti A., Mehdad Y., Bentivogli L., Giampiccolo D. Semeval-2012 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012). 2012.

Turney P.D. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In EMCL 01: Proceedings of the 12th European Conference on Machine Learning, London, UK, Springer-Verlag pages 491502